Trends in Cell Biology

CellPress

Special Issue: Quantitative Cell Biology

Opinion

Theory in Biology: Figure 1 or Figure 7?

Rob Phillips[1,*]

The pace of modern science is staggering. The quantities of data now flowing from DNA sequencers, fluorescence and electron microscopes, mass spectrometers, and other mind-blowing instruments leave us faced with information overload. This explosion in data has brought on its heels a concomitant need for efforts at the kinds of synthesis and unification we see in theoretical physics. Often in cell biology, when theoretical modeling takes place, it is as a figure 7 reflection on experiments that have already been done, with data fitting providing a metric of success. Figure 1 theory, by way of contrast, is about living dangerously by turning our thinking into formal mathematical predictions and confronting that math with experiments that have not yet been done.

What is the Role of Theory in the Life Sciences?

People say that to learn about the philosophy of science, one should not listen to what scientists say, but rather watch what they do. Most of the time, if cell biologists use theory at all, it appears at the end of their paper, a parting shot from figure 7. A model is proposed after the experiments are done, and victory is declared if the model 'fits' the data. But there is another way to go about using theory. This second approach not only provides a conceptual framework for experiments that have already been done but, more importantly, it also uses theory to produce interesting, testable predictions about experiments that have not yet been done. This type of theory often appears at the beginning of the paper, an opening volley from figure 1, to justify the experiments that follow. Here I describe the opportunity offered by practicing 'Figure 1 theory', where the theory comes first, and everything from the experimental design to the data analysis and interpretation flow from it.

It is an important time to reexamine the role of theory in biology. The explosion of data in the life sciences has created a deep tension between fact and concept. Indeed, the frenzy surrounding big data has led some to speculate 'the end of theory' [1]. The supposition is that if we can find the right correlations between different measurables, we need not bother with finding the underlying 'laws' that give rise to those correlations. The French mathematician Henri Poincaré famously noted 'A science is built up of facts as a house is built up of bricks. But a mere accumulation of facts is no more a science than a pile of bricks is a house'. Biology has many rooms and hallways of exquisite beauty, but there are still many bricks awaiting their place in the structure of biological science. Examples abound. Quantitative microscopy is now providing a picture of when and where the macromolecules of the cell are found. Mass spectrometry and fluorescence microscopy give an unprecedented look at the mean and variability in the number of mRNAs, lipids, proteins, and metabolites in cells of all kinds. DNA sequencing now routinely provides a base pair resolution view of genomes and their occupancy by proteins such as histones and transcription factors. Yet we are often lost amid the massive omic and imaging databases we have collected without a theoretical understanding to guide us. When successful,

Trends

The rapid pace of experimental advance and acquisition of exciting news kinds of data in cell biology makes it ever more important to develop conceptual frameworks that unify and explain that data.

Mathematical theory forces us to formally state our thoughts in the same way that writing a computer program demands a precise statement of the underlying algorithm.

Theoretical models complement biochemistry, genetics, bioinformatics, and other frameworks for querying biological systems.

Theory allows us to sharpen our thinking and hypotheses.

[1]Department of Applied Physics and Division of Biology and Biological Engineering, California Institute of Technology, Pasadena, CA, USA

*Correspondence: phillips@pboc.caltech.edu (R. Phillips).

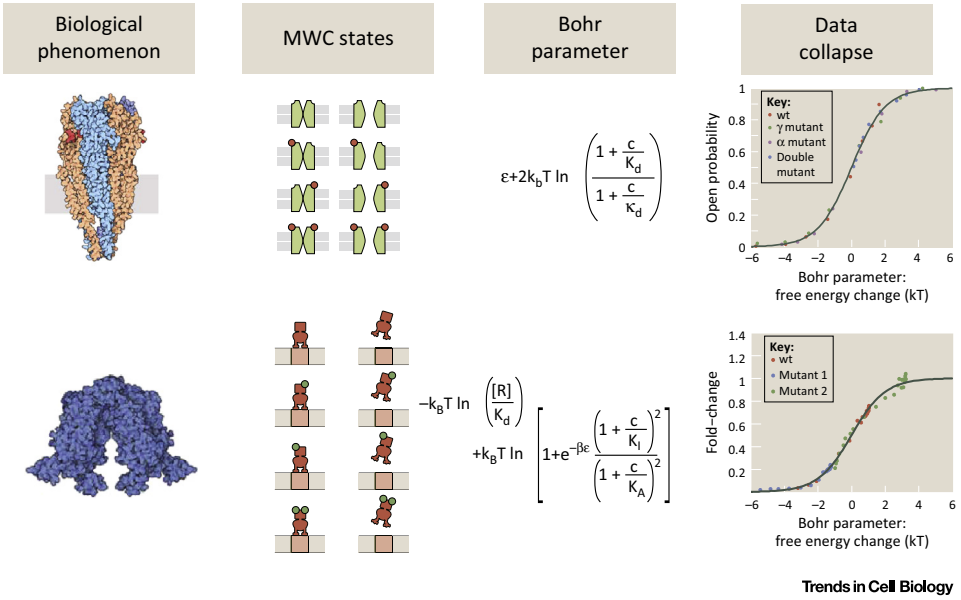

Figure 1. Broad Reach of Statistical Mechanical Models of Allostery. The top example shows an ion channel known as the nicotinic acetylcholine receptor and the bottom example shows the gene regulatory molecule known as Lac repressor. The Monod–Wyman–Changeux model (MWC) considers the inactive and active states in all of their different states of ligand occupancy [36]. The Bohr parameter provides the critical natural scaling variable that makes it possible for data from different mutants to all fall on one master curve as shown in the final column [27]. Different colored data points correspond to different mutants of the ion channel (top) or repressor molecule (bottom). Ion channel data from [37] and repressor data from [38].

Figure 1 theory tells us from the get-go exactly what data we need to collect to attempt to test our theoretical musings. As a result of the experimental advances driving cell biology, there is enormous pressure to turn facts into a corresponding conceptual picture of how cells work [2].

What exactly do we mean by theory? In many cases, our first understanding of some biological problem might be based on powerful, cartoon-level abstractions, already a useful first level of theory that can itself serve a Figure 1 role. These abstractions make qualitative predictions that we can then test. However, by mathematicizing these cartoon-level abstractions, we go farther, by formally committing to their underlying assumptions we can thus use the logical machinery of mathematics to sharpen our hypotheses and more deeply explore their consequences. Jeremy Gunawardena has amusingly but thoughtfully referred to this kind of theory as the exercise of converting our 'pathetic' thinking into mathematical form and then exploring the consequences of the assumptions behind that thinking [3].

How Can Theory Enlighten Us?

Where is the evidence that mathematical theory has the power to expand our understanding of the living world in the same way that microscopy, genetics, and biochemistry, for example, already have? In fact, as has been noted elsewhere, there is a long tradition of deep and fundamental biological insights that required quantitative analysis [3,4]. One of my personal favorites concerns the question of the physical limits on how cells can detect environmental stimuli. Quantitative reasoning has provided us with insights into processes as diverse as chemotaxis, in which cells can detect tiny chemical gradients, or vision, where networks of molecules make it possible for photoreceptors to detect small numbers of photons [5–7]. For example, in the context of chemotaxis, theoretical considerations shed deep light on the mechanisms of both gradient detection and how cells adapt to changes in the ambient chemoattractant concentration [5–8]. Another celebrated example is the way in which probability distributions serve as a window into biological mechanisms [9]. The famed Luria–Delbrück

experiment on the origins of genetic mutations provided critical insights into the mechanisms of evolution across all domains of life [10]. Similarly, the ongoing debate over when the statistics of mRNA distributions are characterized by the Poisson distribution is helping to clarify the mechanistic underpinnings of the processes of the central dogma as they unfold inside a cell [11–14].

When successful, theory can bring us both enlightenment and surprise. One form of enlightenment is through the existence of what one might call metaconcepts. Think about all of the different scenarios in the natural world where the notion of 'resonance' (one of the most far-reaching metaconcepts I can think of) shows up, whether in the back-and-forth motion of a child on a swing or the optical resonators that are central to the many ways we now sculpt light. A compelling biological example is offered by the mathematics of repeated trials of some experiment with only two outcomes (e.g., the familiar heads and tails from a coin flip). This thinking, although seemingly very remote from biology, is actually an overarching theme for understanding many biological processes. For example, thinking about coin flips provides a quantitative basis for answering the question of whether the segregation of carboxysomes in cyanobacteria is an active process or rather the result of random partitioning [15]. Further, for those cases in which molecular partitioning during cell division is random, coin-flip thinking provides a powerful means of converting arbitrary fluorescence units from our microscopy experiments into precise molecular counts [16–19].

One of the most beautiful examples of a metaconcept from biology is provided by the notion of allostery as shown in Figure 1 [20–22]. A wide variety of different biological phenomena are mediated by molecules that can exist in two different conformational states, one that we will dub the active state and the other the inactive state. A crucial feature of these molecules is that they can bind a ligand that has different binding affinities for the active and inactive states, thereby biasing the relative probabilities of these two states. By speaking the language of mathematics, it is possible to unite phenomena as diverse as the Bohr effect in hemoglobin, the accessibility of genomic DNA to DNA-binding proteins, the response of chemotaxis receptors to changes in chemoattractant concentration, the analysis of mutants in quorum sensing, and the induction of transcription factors. As hinted at in Figure 1, all of these phenomena can be described by a single equation that parameterizes their activity as a function of ligand concentration, revealing a deep unity that is hidden from view when these problems are discussed verbally, although many theoretical challenges remain (see Outstanding Questions) [23–27].

As scientists, we are often interested in finding unifying principles. How do we know when we find them? The ability to collapse the results of more than one experiment onto a single master curve reveals that we might have found some deeper concepts that unite apparently distinct phenomena. Stated differently, such data collapses suggest that we have found the natural variables of a given problem. An example of this has already been shown for the case of allostery in Figure 1, where the natural variable is the Bohr parameter. The quantitative study of gene expression provides another attractive example of this idea. The input–output function of a given genetic regulatory architecture depends on the constellation of binding sites for transcription factors that can either activate or repress transcription. For the simplest of these regulatory architectures, namely, simple repression where a single binding site for a repressor controls expression, we define the fold-change in gene expression as the ratio of two quantities, the level of expression in the presence of repressor over the level of expression in its absence. In this case, the relation between fold-change in gene expression and the number of repressors (R) is given by the formula

$$\text{fold-change} = \left(1 + \frac{R}{N_{NS}} e^{-\beta \Delta \varepsilon_{rd}}\right)^{-1}, \quad [1]$$

where N_{NS} is the number of sites in the genome and $\Delta\varepsilon_{rd}$ measures the binding energy of the repressor on its operator [28–31]. By way of contrast, when there are multiple promoters (N)

competing for the attention of those same repressors, the expression for gene expression is given by the intuitively unenlightening equation

$$\text{fold-change} = \frac{1}{N}\frac{\sum_{i=0}^{N}\binom{N}{i}(N-i)\prod_{j=1}^{i}\frac{(2R-j+1)}{N_{NS}}e^{-\beta\Delta\varepsilon_{rd}(2R-j+1)}\theta(2R-j+1)}{\sum_{i=0}^{N}\binom{N}{i}\prod_{j=1}^{i}\frac{(2R-j+1)}{N_{NS}}e^{-\beta\Delta\varepsilon_{rd}(2R-j+1)}\theta(2R-j+1)}, \quad [2]$$

shown here not to convey understanding but rather to reveal obscurity! Part of the reason that this equation is so hard to parse is that it does not reflect the 'natural variables' of the problem [32]. Many biological processes are first formulated in terms of the variables we know the most about. In the case of gene expression this might be the concentration of transcription factors and their affinity for their cognate binding sites. Equations 1 and 2 describe the simple repression regulatory function in terms of these variables and are plotted in Figure 2 (top left). But this is not the most revealing form for these equations. If they are reformulated in terms of an aggregate parameter – the fugacity, λ_r – we can see how different gene regulatory functions are related to each other [33], allowing us to write an equation for the fold-change in gene expression that absorbs both of the previous expressions as

$$\text{fold-change} = \left(1 + \lambda_r e^{-\beta\epsilon_r}\right)^{-1}, \quad [3]$$

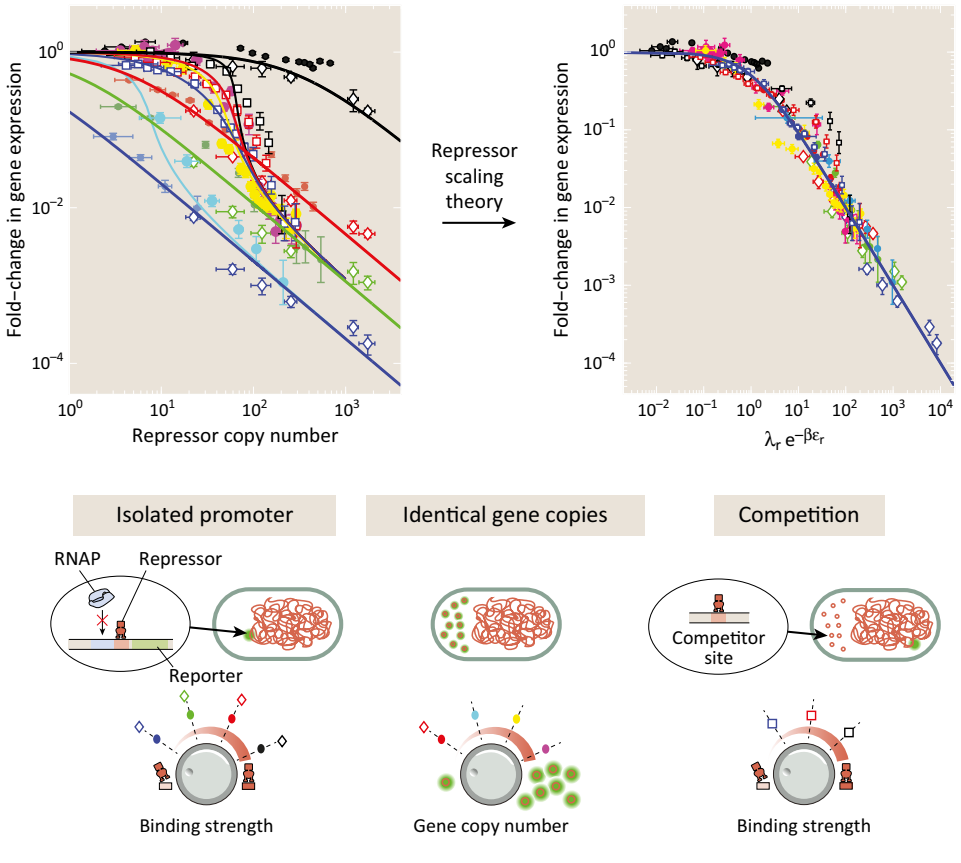

Figure 2. Bringing Different Repression Problems into the Same Fold. The top left graph shows measurements of fold-change in gene expression for a number of different simple repression scenarios including different binding site strengths, repressor copy numbers, and numbers of repressor binding sites across the genome. The data collapse on the top right shows a parameter-free treatment of the same problem in terms of repressor fugacity. The bottom panel shows the different regulatory knobs that are used to control gene expression and that are all accounted for in the fugacity framework [33].

plotted in Figure 2 (top right). Interestingly, the fugacity formulation accounts for three effects simultaneously: (i) transcription factor copy number, (ii) transcription factor binding strength, and (iii) the competition for multiple sites on the DNA for the same transcription factors. One of the exciting outcomes of a theoretical description like this is that it can offer a view in which things that were apparently different are not different at all.

How Can Theory Surprise Us?

Sometimes, people use the word surprise if they find a particular fact to be novel. Further, they might dismiss a theoretical effort by noting some particular theoretical analysis is reasoning about facts that are already known. For example, 'we already know that protein X phosphorylates protein Y resulting in transcription', the implication being that digging into the problem quantitatively offers nothing new or surprising since the key facts are already in hand. I would like to distinguish between finding a fact surprising and finding discrepancies between a conceptual model and data surprising. Each has its own important place in the evolution of understanding of a given phenomenon. To illustrate this, consider trying to predict the tides. That same argument about phosphorylation when turned to the tides would read 'it is not surprising that tides are higher during a full moon', a value judgment based on the primacy of facts over predictive understanding. If you watch the sea and the sun and the moon all day, indeed, you may come to the conclusion that the tides are higher during a full moon. But this is a far cry from the kind of substantive understanding that makes it possible to say how the tides vary every minute of the day, every day of the year, and, further, how those tides vary from one point on the California coast to another. Overall, the resulting theory that allows us to predict tides still tells you that the tides are higher during a full moon, so are you not surprised because you already knew that fact? In my opinion, it is often when we turn the current best understanding of a given biological problem into mathematical language and use it to make quantitative predictions that we are then able to know what is surprising and what is not.

A biological example that makes this point is illustrated in Figure 3. One of the most intriguing aspects of genomes is action at a distance, referring to the fact that binding of proteins on one part of the genome can affect what happens elsewhere on the genome. Perhaps the most well-known example of this kind of effect is the presence of enhancers in the genomes of multicellular organisms. But even bacteria exhibit action at a distance with transcription factors binding at several sites simultaneously and looping the intervening DNA as shown schematically in Figure 3. In the results of this now classic experiment (see Figure 3C), the level of gene expression was measured as a function of the distance between two repressor binding sites [34]. Is the curve shown in Figure 3 surprising? Several features that we can wonder about are the periodicity of

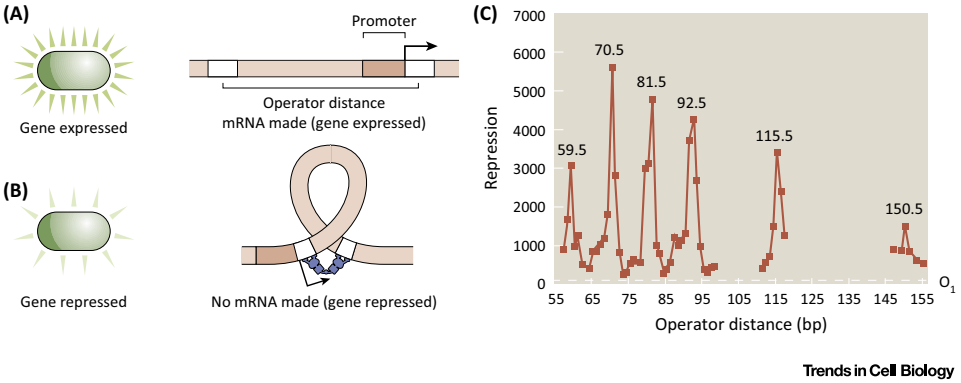

Figure 3. Repression as a Function of Loop Length. (A) Promoter not occupied by repressor, gene expression is on. (B) Promoter occupied by repressor, forming a DNA loop and turning off gene expression. (C) Gene expression as a function of distance between the two repressor binding sites [34]. Is this graph surprising?

the graph as well as the amplitudes of the peaks. Although the periodicity can be attributed to the approximately 10 base pair helical repeat of the DNA double helix, the amplitudes of the peaks and especially the maximum at 70.5 base pairs remains a deep and surprising problem, but only when viewed through the quantitative filter of what we know about DNA elasticity, which tells us that at length scales shorter than the persistence length the DNA should be difficult to bend.

A Role for Figure 1 Theory

One explanation for indifference to examples such as that of DNA looping as deep theoretical challenges is what I would liken to the Kazakh poetry effect. The argument goes something like this: there is no first rate poetry from Kazakhstan. How might someone come to this view? Well, because how many of us have actually seen or heard an outstanding Kazakh poem? The fallacy in this thinking can be carried over to the biological context. It is hard to appreciate the beauty of a language that one does not even speak so when a thoughtful and interesting biological argument is made in mathematical language, there is a risk that some people will not understand it. Of course, this cuts in both ways. Just as it might be hard for those that do not use mathematics as their natural language for describing the world to find a given mathematically reasoned hypothesis surprising, there are many languages within biology (e.g., genetics, bioinformatics, etc.) that are similarly opaque to those who do not speak them. There is no shortage of physical scientists who are ready to make misguided and dogmatic pronouncements about the lack of rigor in biology, or its supposed lack of fascinating problems and deep concepts.

In a recent book, noted historian of science Stephen Brush uses the history of physics, chemistry, and biology to explore the circumstances under which new theories are adopted [35]. He notes that the scientific method as represented in oversimplified textbooks argues that 'adoption' of new theories is supposedly always based on predicting the results of experiments that have not yet been done, perhaps best exemplified by the way Mendeleev's periodic table heralded the existence of new chemical elements. This view is founded on the idea that it is only when those concepts predict something that has not been known before that concepts are truly adopted. Interestingly, Brush argues that in many famous cases from physics such as the adoption of both general relativity and quantum mechanics, that instead, it was the explanation of already known effects that carried the most weight. Although Brush's observations make it clear that figure 7 theory has its place, in my view, there has thus far been a largely missed opportunity to use Figure 1 theory as a guide to sharpen our thinking and to help us design experiments that otherwise would not even have been thought of. Living organisms exhibit beautiful and surprising phenomena at every turn. In my view, there is no one approach that guarantees success in uncovering the secrets of the living world. The thesis of this brief essay is that theoretical descriptions of biological phenomena couched in the language of mathematics have the capacity of revealing insights that would otherwise remain hidden. Future directions for these approaches are presented in Outstanding Questions.

Outstanding Questions

To what extent is biology amenable to the kind of rich interplay between theory and experiment more familiar in its more quantitative partner sciences?

Can biological measurements get to the point of precision and reproducibility that they suffice to distinguish between competing quantitative models?

How do we construct conceptual frameworks that allow us to tame the 'big data' that has become a mainstay of modern biology?

How can we construct a theoretical understanding of the data now flowing from DNA sequencers, fluorescence and electron microscopes, mass spectrometers, and other impressive modern instruments?

Acknowledgments

I dedicate this brief essay to Eric Davidson, friend and colleague, who always insisted on concepts over facts. I am grateful to Stephanie Barnes, Nathan Belliveau, Justin Bois, Griffin Chure, Angela DePace, Tal Einav, Hernan Garcia, Jeremy Gunawardena, Soichi Hirokawa, Greg Huber, Jane Kondev, Madhav Mani, Ron Milo, Muir Morrison, Andrew Murray, Arjun Raj, Manuel Razo, Allyson Sgro, and Jacob Shenker for useful discussions. I am privileged to be entrusted by the National Science Foundation, the National Institutes of Health (NIH), The California Institute of Technology, and La Fondation Pierre Gilles de Gennes with the funds that make the kind of theoretical work described here possible. Specifically, I am grateful to the NIH for support through award numbers DP1 OD000217 (Directors Pioneer Award) and R01 GM085286. I am also grateful to the Kavli Institute for Theoretical Physics where this essay was written, for their generous support. Finally, the referencing provided here is cursory and intended to guide the reader to the literature and does not attempt to provide a scholarly assessment of the many excellent contributions not cited here.

References

1. Anderson, C. (2008) The end of theory: the data deluge makes the scientific method obsolete. *Wired* 16, July (http://www.wired.com/2008/06/pb-theory/)
2. Bray, D. (2001) Reasoning for results. *Nature* 412, 863
3. Gunawardena, J. (2014) Models in biology: 'accurate descriptions of our pathetic thinking'. *BMC Biol.* 12, 29
4. Phillips, R. and Milo, R. (2009) A feeling for the numbers in biology. *Proc. Natl. Acad. Sci. U.S.A.* 106, 21465–21471
5. Berg, H.C. and Purcell, E.M. (1977) Physics of chemoreception. *Biophys. J.* 20, 193–219
6. Phillips, R. *et al.* (2013) *Physical Biology of the Cell.* (2nd edn), Garland Science
7. Bialek, W. (2012) *Biophysics: Searching for Principles,* Princeton University Press
8. Barkai, N. and Leibler, S. (1997) Robustness in simple biochemical networks. *Nature* 387, 913–917
9. Frank, S.A. (2014) How to read probability distributions as statements about process. *Entropy* 16, 6059–6098
10. Luria, S.E. and Delbruck, M. (1943) Mutations of bacteria from virus sensitivity to virus resistance. *Genetics* 28, 491–511
11. Zenklusen, D. *et al.* (2008) Single-RNA counting reveals alternative modes of gene expression in yeast. *Nat. Struct. Mol. Biol.* 15, 1263–1271
12. Taniguchi, Y. *et al.* (2010) Quantifying *E. coli* proteome and transcriptome with single-molecule sensitivity in single cells. *Science* 329, 533
13. So, L.H. *et al.* (2011) General properties of transcriptional time series in *Escherichia coli*. *Nat. Genet.* 43, 554–560
14. Sanchez, A. and Golding, I. (2013) Genetic determinants and cellular constraints in noisy gene expression. *Science* 342, 1188–1193
15. Savage, D.F. *et al.* (2010) Spatially ordered dynamics of the bacterial carbon fixation machinery. *Science* 327, 1258–1261
16. Golding, I. *et al.* (2005) Real-time kinetics of gene activity in individual bacteria. *Cell* 123, 1025–1036
17. Rosenfeld, N. *et al.* (2005) Gene regulation at the single-cell level. *Science* 307, 1962–1965
18. Teng, S.W. *et al.* (2010) Measurement of the copy number of the master quorum-sensing regulator of a bacterial cell. *Biophys. J.* 98, 2024–2031
19. Brewster, R.C. *et al.* (2014) The transcription factor titration effect dictates level of gene expression. *Cell* 156, 1312–1323
20. Changeux, J.P. (2013) 50 years of allosteric interactions: the twists and turns of the models. *Nat. Rev. Mol. Cell Biol.* 14, 819–829
21. Martins, B.M. and Swain, P.S. (2011) Trade-offs and constraints in allosteric sensing. *PLoS Comput. Biol.* 7, e1002261
22. Marzen, S. *et al.* (2013) Statistical mechanics of Monod–Wyman–Changeux (MWC) models. *J. Mol. Biol.* 425, 1433–1460
23. Mello, B.A. and Tu, Y. (2003) Quantitative modeling of sensitivity in bacterial chemotaxis: the role of coupling among different chemoreceptor species. *Proc. Natl. Acad. Sci. U.S.A.* 100, 8223–8228
24. Keymer, J.E. *et al.* (2006) Chemosensing in *Escherichia coli*: two regimes of two-state receptors. *Proc. Natl. Acad. Sci. U.S.A.* 103, 1786–1791
25. Swem, L.R. *et al.* (2008) Deducing receptor signaling parameters from in vivo analysis: LuxN/AI-1 quorum sensing in *Vibrio harveyi*. *Cell* 134, 461–473
26. Mirny, L.A. (2010) Nucleosome-mediated cooperativity between transcription factors. *Proc. Natl. Acad. Sci. U.S.A.* 107, 22534–22539
27. Phillips, R. (2015) Napoleon is in equilibrium. *Annu. Rev. Condens. Matter Phys.* 6, 85–111
28. Buchler, N.E. *et al.* (2003) On schemes of combinatorial transcription logic. *Proc. Natl. Acad. Sci. U.S.A.* 100, 5136–5141
29. Bintu, L. (2005) Transcriptional regulation by the numbers: models. *Curr. Opin. Genet. Dev.* 15, 116–124
30. Bintu, L. *et al.* (2005) Transcriptional regulation by the numbers: applications. *Curr. Opin. Genet. Dev.* 15, 125–135
31. Garcia, H.G. and Phillips, R. (2011) Quantitative dissection of the simple repression input–output function. *Proc. Natl. Acad. Sci. U.S.A.* 108, 12173–12178
32. Rydenfelt, M. *et al.* (2014) Statistical mechanical model of coupled transcription from multiple promoters due to transcription factor titration. *Phys. Rev. E* 89, 012702
33. Weinert, F.M. *et al.* (2014) Scaling of gene expression with transcription-factor fugacity. *Phys. Rev. Lett.* 113, 258101
34. Muller, J. *et al.* (1996) Repression of *lac* promoter as a function of distance, phase and quality of an auxiliary *lac* operator. *J. Mol. Biol.* 257, 21–29
35. Brush, S.G. (2015) *Making 20th Century Science: How Theories Become Knowledge,* Oxford University Press
36. Monod, J. *et al.* (1965) On the nature of allosteric transitions: a plausible model. *J. Mol. Biol.* 12, 88–118
37. Labarca, C. *et al.* (1995) Channel gating governed symmetrically by conserved leucine residues in the M2 domain of nicotinic receptors. *Nature* 376, 514–516
38. Daber, R. *et al.* (2011) Thermodynamic analysis of mutant lac repressors. *J. Mol. Biol.* 409, 76–87